Manuscript Title

This manuscript (permalink) was automatically generated from greenelab/generalization-manuscript@584094e on July 13, 2023.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe · @johndoe@mastodon.social
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

Methods

Mutation data download and preprocessing

To generate binary mutated/non-mutated gene labels for our machine learning model, we used mutation calls for TCGA samples from MC3 [1] and copy number threshold calls from GISTIC2.0 [2]. MC3 mutation calls were downloaded from the Genomic Data Commons (GDC) of the National Cancer Institute, at https://gdc.cancer.gov/about-data/publications/pancanatlas. Thresholded copy number calls are from an older version of the GDC data and are available here: https://figshare.com/articles/dataset/TCGA_PanCanAtlas_Copy_Number_Data/6144122. We removed hypermutated samples, defined as two or more standard deviations above the mean non-silent somatic mutation count, from our dataset to reduce the number of false positives (i.e., non-driver mutations). Any sample with either a non-silent somatic variant or a copy number variation (copy number gain in the target gene for oncogenes and copy number loss in the target gene for tumor suppressor genes) was included in the positive set; all remaining samples were considered negative for mutation in the target gene.

We followed a similar procedure to generate binary labels for cell lines from CCLE, using the data available on the DepMap download portal at https://depmap.org/portal/download/all/. Mutation information was retrieved from the `OmicsSomaticMutations.csv` data file, and copy number inforamtion was retrieved from the `OmicsCNGene.csv` data file. We thresholded the CNV log-ratios provided by CCLE into binary gain/loss calls using a lower threshold of $\log_2(3/2)$ (i.e. cell lines with a log-ratio below this threshold were considered to have a full copy loss in the corresponding gene), and an upper threshold of $\log_2(5/2)$ (i.e. cell lines with a log-ratio above this threshold were considered to have a full copy gain in the corresponding gene). After applying the same hypermutation criteria that we used for TCGA, no cell lines in CCLE were identified as hypermutated. After preprocesing, 1402 cell lines with mutation and copy number data remained. We then combined non-silent point mutations and copy number gain/loss information into binary labels using the same criteria as for TCGA.

Gene expression data download and preprocessing

RNA sequencing data for TCGA was downloaded from GDC at the same link provided above for the Pan-Cancer Atlas. We discarded non-protein-coding genes and genes that failed to map, and removed tumors that were measured from multiple sites. After filtering to remove hypermutated samples and taking the intersection of samples with both mutation and gene expression data, 9074 TCGA samples remained.

RNA sequencing data for CCLE was downloaded from the DepMap download portal, linked above, in the `CCLE_expression.csv` data file. After taking the intersection of CCLE cell lines with both mutation and gene expression data, 1402 cell lines remained. For experiments making predictions across datasets (i.e., training models on TCGA and evaluating performance on CCLE, or vice-versa) we took the intersection of genes in both datasets, resulting in 16041 gene features. For experiments where only TCGA data was used (i.e., evaluating models on held-out cancer types), we used all 16148 gene features present in TCGA after the filtering described above.

Cancer gene set construction

In order to study mutation status classification for a diverse set of cancer driver genes, we started with the set of 125 frequently altered genes from Vogelstein et al. [3] (all genes from Table S2A). For each target gene, to ensure that the training dataset was reasonably balanced (i.e., that there would

be enough mutated samples to train an effective classifier), we included only cancer types with at least 15 mutated samples and at least 5% mutated samples, which we refer to here as “valid” cancer types. In some cases, this resulted in genes with no valid cancer types, which we dropped from the analysis. Out of the 125 genes originally listed in the Vogelstein et al. cancer gene set, we retained 71 target genes for the TCGA to CCLE analysis, and 70 genes for the CCLE to TCGA analyses. For these analyses, each gene needed at least one valid cancer type in TCGA and one valid cancer type in CCLE, to construct the train and test sets. For the cancer type holdout analysis, we retained 56 target genes: in this case, each gene needed at least two valid cancer types in TCGA to be retained, one to train on and one to hold out.

Classifier setup and cross-validation design

LASSO logistic regression details LASSO range selection Neural network details? (if necessary)

“Best model” vs. “smallest good model” analysis details

Description of “smallest good” heuristic Statistical testing?

Open science/reproducibility stuff

Results

"Best" and "smallest good" model selection strategies perform comparably

To address the question of whether more parsimonious models tend to generalize better or not, we designed two model selection schemes and compared them for the TCGA to CCLE and CCLE to TCGA mutation prediction problems (Figure 1A). The "best" model selection scheme chooses the top-performing model/LASSO parameter on the holdout dataset from the same source as the training data and applies it to the test data from the other data source. The intention of the "smallest good" model selection scheme is to balance parsimony with reasonable performance on the holdout data, since simply selecting the smallest possible model (generally, the dummy regressor/mean predictor) is not likely to generalize well. To accomplish this, we first identify the top 25% of well-performing models on the holdout dataset; then, from this subset of models, we choose the smallest (i.e., highest LASSO parameter) to apply to the test data. In both cases, we exclusively use the holdout data to select a model and only apply the model to out-of-dataset samples to evaluate generalization performance *after* model selection.

For TCGA to CCLE generalization, 27/71 genes (38.0%) had better performance for the "best" model, and 17/71 genes (23.9%) had better generalization performance with the "smallest good" model. The other 27 genes had the same "best" and "smallest good" model (in other words, the "smallest good" model was also the best-performing overall, and the difference was 0) (Figure 1B). For CCLE to TCGA generalization, 23/70 genes (32.9%) had better performance for the "best" model and 18/70 (25.7%) for the "smallest good," with the other 29 having the same model fulfill both criteria (Figure 1C). Overall, these results do not support the hypothesis that the most parsimonious model generalizes the best: for both generalization problems there are slightly more genes where the best-performing model on the holdout dataset is also the best-performing on the test set, although there are some genes where the "smallest good" approach works well.

We examined genes that fell into either category for TCGA to CCLE generalization (dotted lines on Figure 1B). For *NF1*, the "best" model outperforms the "smallest good" model (Figure 1D). Comparing holdout (orange) and cross-dataset (green) performance, both generally follow a similar trend, with the cross-dataset performance peaking when the holdout performance peaks at a regularization parameter of $\alpha = 0.00316$. *PIK3CA* is an example of the opposite, a gene where the "smallest good" model tends to outperform the "best" model (Figure 1E). In this case, the peak for the cross-dataset performance occurs at a higher level of regularization (further left on the x-axis), at $\alpha = 0.01$, than the peak for the holdout performance, at $\alpha = 0.0316$. This suggests that a *PIK3CA* mutation status classifier that is more parsimonious, but that has slightly worse performance, does tend to generalize better across datasets to CCLE.

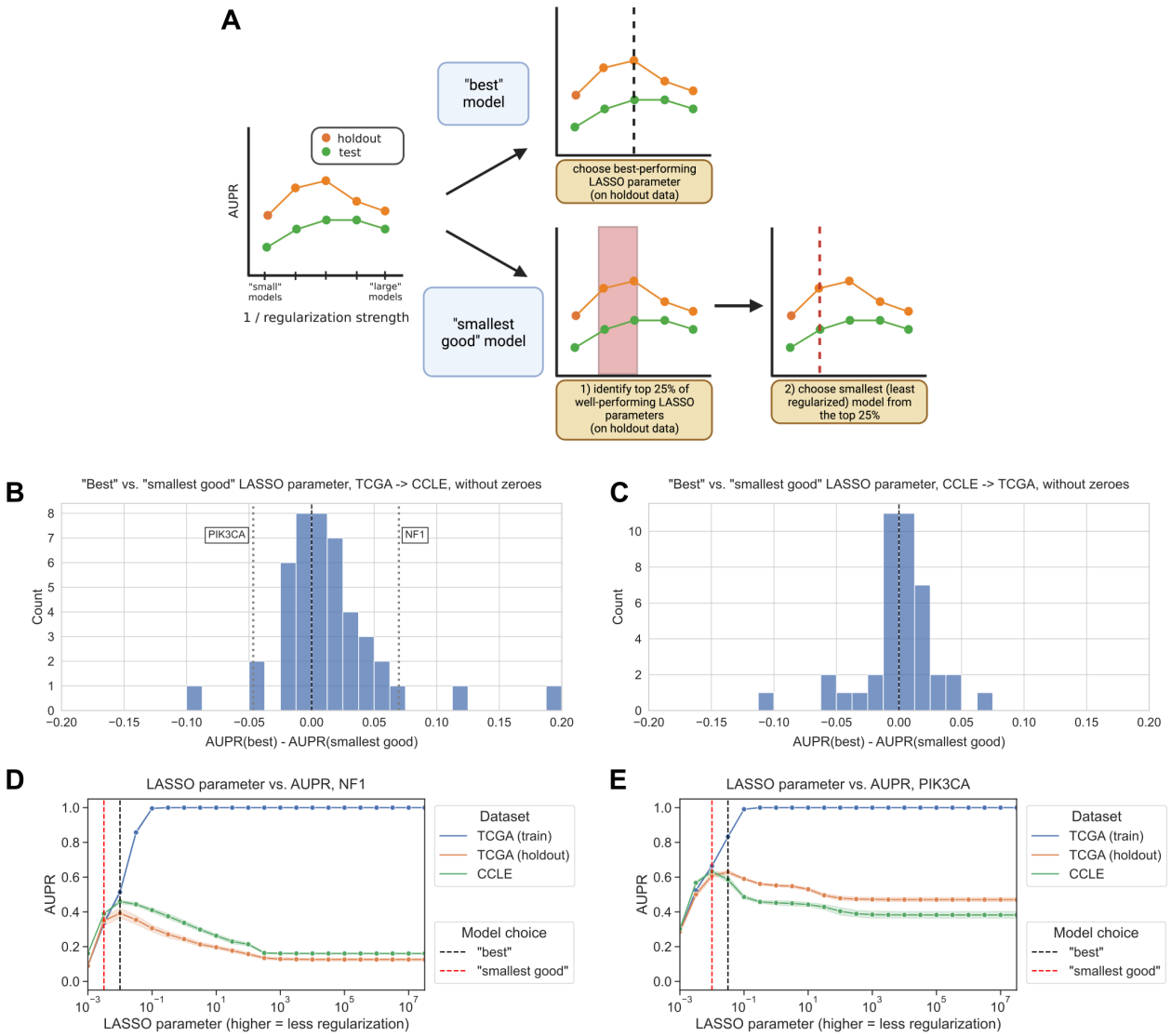

Figure 1: **A.** Schematic of "best" vs. "smallest good" model comparison experiments. **B.** Distribution of performance comparisons between "best" and "smallest good" model selection strategies, for TCGA -> CCLE generalization. Positive x-axis values indicate better performance for the "best" model, negative values indicate better performance for the "smallest good" model. **C.** Distribution of performance comparisons between "best" and "smallest good" model selection strategies, for CCLE -> TCGA generalization. **D.** *NF1* mutation status prediction performance generalizing from TCGA (holdout, orange), to CCLE (green), with "best" and "smallest good" models labeled. **E.** *PIK3CA* mutation status prediction performance generalizing from TCGA (holdout, orange), to CCLE (green), with "best" and "smallest good" models labeled.

References

1. **Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines**
   Kyle Ellrott, Matthew H Bailey, Gordon Saksena, Kyle R Covington, Cyriac Kandoth, Chip Stewart, Julian Hess, Singer Ma, Kami E Chiotti, Michael McLellan, … Armaz Mariamidze
   Cell Systems (2018-03) https://doi.org/gf9twn
   DOI: 10.1016/j.cels.2018.03.002 · PMID: 29596782 · PMCID: PMC6075717

2. **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers**
   Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, Gad Getz
   Genome Biology (2011-04) https://doi.org/dzhjqh
   DOI: 10.1186/gb-2011-12-4-r41 · PMID: 21527027 · PMCID: PMC3218867

3. **Evaluating the evaluation of cancer driver genes**
   Collin J Tokheim, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein, Rachel Karchin
   Proceedings of the National Academy of Sciences (2016-11-22) https://doi.org/f9d77w
   DOI: 10.1073/pnas.1616440113 · PMID: 27911828 · PMCID: PMC5167163